

Genomic organisation of the seven ParaHox genes of coelacanth

Mulley, J.F.; Holland, P.W.

Journal of Experimental Zoology Part B: Molecular and Developmental Evolution

DOI:

[10.1002/jez.b.22513](https://doi.org/10.1002/jez.b.22513)

Published: 01/09/2014

Publisher's PDF, also known as Version of record

[Cyswllt i'r cyhoeddiad / Link to publication](#)

Dyfyniad o'r fersiwn a gyhoeddwyd / Citation for published version (APA):

Mulley, J. F., & Holland, P. W. (2014). Genomic organisation of the seven ParaHox genes of coelacanth. *Journal of Experimental Zoology Part B: Molecular and Developmental Evolution*, 322(6), 352-358. <https://doi.org/10.1002/jez.b.22513>

Hawliau Cyffredinol / General rights

Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal ?

Take down policy

If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Genomic Organisation of the Seven ParaHox Genes of Coelacanth

JOHN F. MULLEY^{1*} AND
PETER W.H. HOLLAND²

¹School of Biological Sciences, Bangor University, Bangor, Gwynedd, United Kingdom

²Department of Zoology, University of Oxford, Oxford, United Kingdom



ABSTRACT

Human and mouse genomes contain six ParaHox genes implicated in gut and neural patterning. In coelacanths and cartilaginous fish, an additional ParaHox gene exists—*Pdx2*—that dates back to the genome duplications in early vertebrate evolution. Here we examine the genomic arrangement and flanking genes of all ParaHox genes in coelacanths, to determine the full complement of these genes. We find that coelacanths have seven ParaHox genes in total, in four chromosomal locations, revealing that five gene losses occurred soon after vertebrate genome duplication. Comparison of intergenic sequences reveals that some *Pdx1* regulatory regions associated with development of pancreatic islets are older than tetrapods, that *Pdx1* and *Pdx2* share few if any conserved non-coding elements, and that there is very high sequence conservation between coelacanth species. *J. Exp. Zool. (Mol. Dev. Evol.)* 322B:352–358, 2014. © 2013 Wiley Periodicals, Inc.

J. Exp. Zool.
(*Mol. Dev. Evol.*)
322B:352–358,
2014

How to cite this article: Mulley JF, Holland PWH. 2014. Genomic organisation of the seven ParaHox genes of coelacanths. *J. Exp. Zool. (Mol. Dev. Evol.)* 322B:352–358.

The ParaHox gene cluster, first described in the cephalochordate amphioxus (Brooke et al., '98), contains three tandemly arrayed homeobox genes (*Gsx*, *Xlox*, and *Cdx*) and dates to the base of the Bilateria (Ferrier and Minguillón, 2003) or earlier (Hui et al., 2008; Mendivil Ramos et al., 2012). The gene cluster has undergone gene losses, rearrangements, and duplications in different animal lineages. The ParaHox genes of bilaterian animals are involved in several developmental processes, notably in gut and neural patterning. Thus, endodermal midgut expression has been described for *Xlox* genes in vertebrates, amphioxus, echinoderm, annelids, and molluscs, with functional importance in gut formation demonstrated for sea urchin (Arnone et al., 2006) and for the well-studied vertebrate *Xlox* gene *Pdx1* (Jonsson et al., '94; Offield et al., '96; Brooke et al., '98). The *Cdx* (*caudal-type homeobox*) genes are expressed posteriorly in many animals, and are important for development of the posterior gut and anus, as well as the posterior neural tube in insects and vertebrates (Moreno and Morata, '99; Reece-Hoyes et al., 2002). *Gsx* (*genomic screen homeobox*, also known as *Gsh* or *ind*) genes are primarily involved in brain rather than gut patterning in mammals (Hsieh-Li et al., '95; Valerius et al., '95), although expression has been reported in the endocrine pancreas (Rosanas-Urgell et al., 2005). In at least some molluscs and annelids, the *Gsx* gene is expressed in the mouth region in addition to some nerve cells (Kulakova et al., 2008; Hui et al., 2009; Samadi and Steiner, 2010), consistent

with the hypothesis that bilaterian ParaHox genes may have ancestrally patterned mouth, midgut, and anus (Holland, 2001).

Genome duplication events early in vertebrate evolution resulted in an increase in the number of ParaHox genes, such that the human genome has six ParaHox genes at four chromosomal locations. A cluster of three genes (*GSX1*, *PDX1/IPF1*, *CDX2*) is located at 13q12.1, but there are no other linked pairs or clusters of ParaHox genes in humans. Instead, three individual genes (*GSX2*, *CDX1*, *CDX4*) are located at three different genomic locations; analysis of flanking genes shows that these are remnants from three degenerate ParaHox gene clusters (Fig. 1). Humans have, therefore, lost six putative ParaHox genes in evolution, but the timing of this loss has been unclear.

Grant sponsor: Wellcome Trust; grant number: 081233/Z/06/Z.

Conflicts of interest: None.

*Correspondence to: John F. Mulley, School of Biological Sciences, Bangor University, Brambell Building, Deiniol Road, Bangor, Gwynedd LL57 2UW, United Kingdom.

E-mail: j.mulley@bangor.ac.uk

Received 27 February 2013; Revised 25 April 2013; Accepted 25 April 2013

DOI: 10.1002/jez.b.22513

Published online 17 June 2013 in Wiley Online Library (wileyonlinelibrary.com).

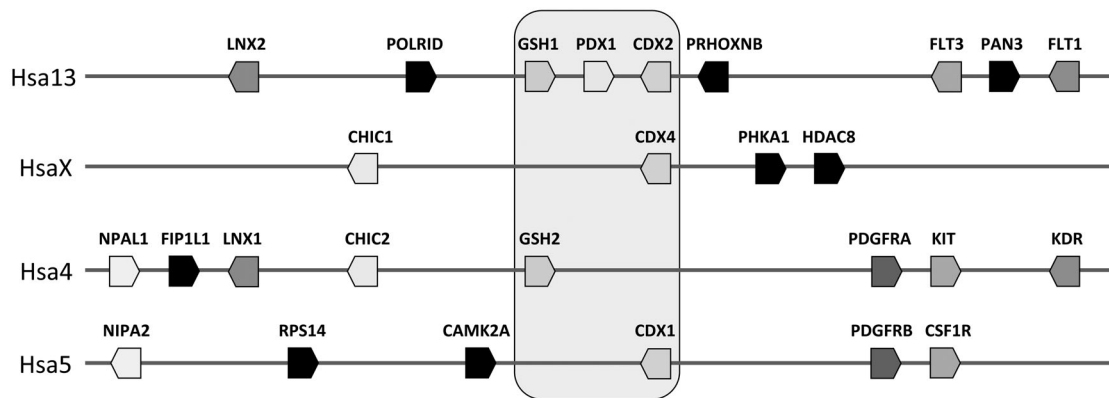


Figure 1. Organisation of ParaHox gene-containing regions in the human genome. An intact cluster of *Gsx*, *Pdx*, and *Cdx* genes is located on chromosome 13 (cluster A), with other ParaHox genes on the X chromosome (cluster B), chromosome 4 (cluster C), and chromosome 5 (cluster D).

A similar complement of ParaHox genes has been found in other jawed vertebrates examined, with three exceptions identified to date. First, zebrafish and pufferfish have lost *cdx2*, but have duplicates of *cdx1* (Mulley et al., 2006). The two *cdx1* genes in teleost fish are a remnant of the additional genome duplication that occurred in teleost fish evolution, though intriguingly this implies that duplicates of (at least) *gsx1*, *gsx2*, *pdx1*, and *cdx4* were lost, alongside two *cdx2* copies. Second, an ancient *Pdx2* gene was recently found in chondrichthyan fish (dogfish and skate), genomically linked to *Gsx2*; this implies that *Pdx2* was present in the common ancestor of extant jawed vertebrates (Mulley and Holland, 2010). Third, and most surprisingly, a *Pdx2* gene was also discovered in the coelacanths, lobe-finned fish that lie within the sarcopterygian clade of vertebrates, close to tetrapods (Mulley and Holland, 2010). The implication is that *Pdx2* existed for much of vertebrate evolution, at least up to the radiation of the sarcopterygians, and was lost independently in ray-finned fish and tetrapods.

The identification of this additional ParaHox gene, *Pdx2*, in coelacanths raised the question of whether other ParaHox genes might have survived for a significant length of time in vertebrate evolution, to be lost later in tetrapods. Coelacanths are sometimes referred to as “living fossils,” meaning that they have retained a relatively unchanged body plan for hundreds of millions of years, and it would be intriguing if their genome also retained ancient genes (such as *Pdx2*) lost in other lineages (Noonan et al., 2004; Powers and Amemiya, 2004; Amemiya et al., 2010). To investigate this, we sequenced and assembled the genomic regions flanking all ParaHox genes in a coelacanth, and their bordering non-homeobox genes, to define the limits of each ParaHox gene cluster or gene cluster remnant. We find a surprising level of sequence conservation between the extant coelacanth species and demonstrate the absence of any additional ParaHox genes beyond the

known seven. We also show that known regulatory regions associated with the *Pdx1* gene are older than previously thought and that the paralogous *Pdx1* and *Pdx2* genes of coelacanths do not appear to share regulatory architecture.

RESULTS

We previously reported that coelacanths have retained an additional ParaHox gene (*Pdx2*) that has been independently lost in other bony vertebrate lineages (Mulley and Holland, 2010) and set out to discover whether they may also have retained additional members of the *Gsx*, *Xlox* and *Cdx* gene families. Sequencing and assembly of two overlapping BAC clones from Indonesian coelacanth yielded a 217,497 bp contiguous sequence containing the full coding sequences of the ParaHox genes *Gsx1*, *Pdx1*, and *Cdx2*, plus complete or partial sequences of the flanking *Prhoxnb* (*ParaHox cluster neighbour*) and *Flt3* (*FMS-like tyrosine kinase 3*) genes (Fig. 2). A 154,768 bp *Pdx2*-containing Indonesian coelacanth BAC clone has been described previously (Mulley and Holland, 2010) and also contains the *Pdgfra* (*Alpha-type platelet-derived growth factor receptor*) gene. A fourth ParaHox-containing BAC clone of 185,347 bp assembled into six ordered pieces and contained the *Cdx4* and *Lnx3* (*Ligand of numb protein X 3*) genes. It has been suggested that the ratio of the relative spacing of genes (*Gsx* to *Xlox*, and *Xlox* to *Cdx*) is conserved in chordate ParaHox clusters (Ferrier et al., 2005); we find that this pattern is broadly conserved in coelacanths (Fig. 3) (see Table S1, Supplementary information).

Evidence for ParaHox genes and their genomic neighbours from these sequenced BAC clones, as well as extensive degenerate PCR experiments (using a variety of primer combinations and reaction conditions), plus analysis of preliminary whole genome sequence information available for the African coelacanth *Latimeria chalumnae*, demonstrates that the complement of ParaHox genes

in coelacanths is: two members of the *Gsx* gene family (*Gsx1*, *Gsx2*); two members of the *Xlox* gene family (*Pdx1* and *Pdx2*) and three members of the *Cdx* gene family (*Cdx1*, *Cdx2*, *Cdx4*) (Fig. 2). We found no evidence for additional ParaHox genes in coelacanths and PCR-based analyses have so far failed to identify any others in other vertebrate groups, including cartilaginous fish (Mulley and Holland, 2010) and lungfish (*Neoceratodus forsteri*, data not shown). We suggest that the situation in coelacanths is representative of the basal jawed vertebrate condition.

The availability of sequence information for the canonical ParaHox A cluster (*Gsx1*, *Pdx1*, *Cdx2*) and the *Gsx2*–*Pdx2* cluster (cluster C) enables comparisons of the conserved non-coding (putative regulatory) sequences associated with the paralogous *Pdx* genes—something that is desirable given the important role played by the *Pdx1* gene in pancreas development (Jonsson et al., '94; Offield et al., '96; Stoffers et al., '97; Schwitzgebel et al., 2003) and in glucose-responsive regulation of insulin gene expression (Ohlsson et al., '93), and its relevance to diabetes (Cockburn et al., 2004; Hani et al., '99; Macfarlane et al., '99). These comparisons were not previously possible with the fragmentary data from earlier studies and shed light on the evolutionary history of pancreas development and insulin regulation.

We first used mVISTA (Frazer et al., 2004) to compare the human, Bowfin (*Amia calva*—a basal ray-finned fish with an

intact cluster (Mulley et al., 2006)) and African and Indonesian coelacanth ParaHox A clusters to identify conserved non-coding regions of DNA (Fig. 4). This revealed a pattern of conservation within and around the canonical ParaHox cluster similar to that in other vertebrates (Mulley et al., 2006). In contrast, no conserved non-coding elements were identified between the Coelacanth *Pdx1* and *Pdx2* regions, apart from a short region with no similarity to known *Pdx1* enhancer sequences and with high sequence conservation to other coelacanth BAC clones in Genbank, suggesting a repetitive element in the Coelacanth genome. We find no conserved non-coding elements between the *Pdx2*-containing regions of *L. chalumnae* and Little Skate (*Leucoraja erinacea*); the most highly conserved elements are close to *Gsx2* and also conserved in the human genome (which lacks *Pdx2*), suggesting they regulate *Gsx2*. To assess if the evolutionary rates of *Pdx1* and *Pdx2* were unequal, we conducted relative rate tests. No significant differences were detected (*L. chalumnae* $P = 0.76808$; *Scyliorhinus canicula* $P = 0.54649$).

The human and rodent *Pdx1* genes have been intensively studied and a number of regions important for transcriptional regulation have been identified, some of which are conserved in other tetrapods (Gerrish et al., 2000, 2004; Boyer et al., 2006; Fujitani et al., 2006; Miyatsuka et al., 2007; Wiebe et al., 2007). We analyzed DNA sequence 5' of the coelacanth *Pdx1* coding regions, searching for evidence of Areas I–III (involved in differentiation

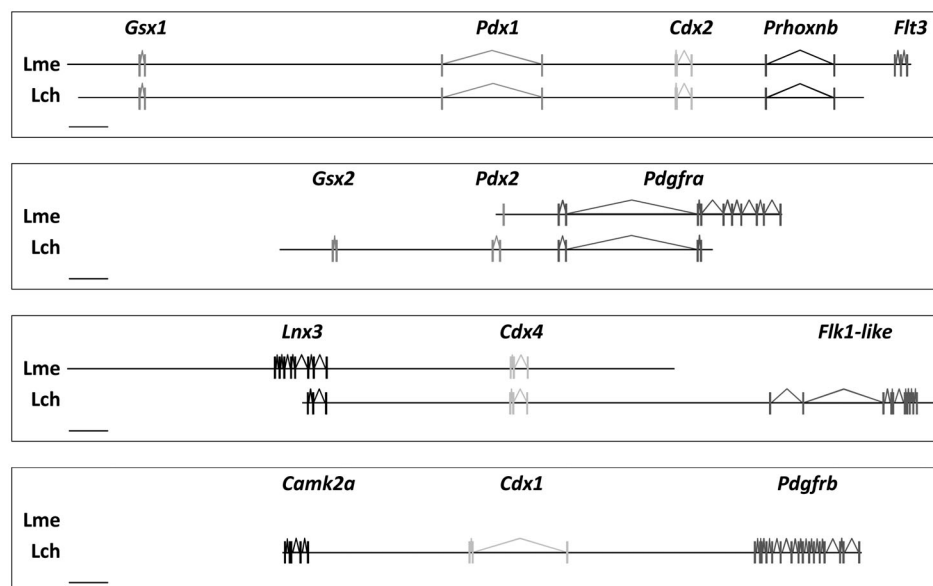


Figure 2. Organisation of ParaHox genes in the two extant species of Coelacanth (Lme—Indonesian coelacanth *Latimeria menadoensis*, Lch—African coelacanth *Latimeria chalumnae*). Lme data is based on sequenced BAC clones and Lch data is derived from genome sequence information. The complement of ParaHox genes in both species is two *Gsx* (*Gsx1*, *Gsx2*), two *Xlox* (*Pdx1*, *Pdx2*), and three *Cdx* (*Cdx1*, *Cdx2*, *Cdx4*). Scale bar is 10 kb.

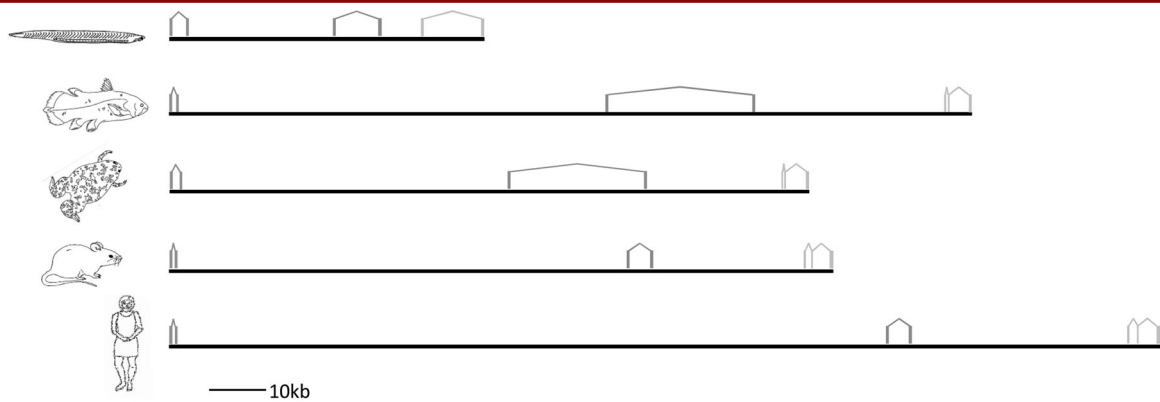


Figure 3. Genomic organisation of the ParaHox A cluster in vertebrates. Vertebrate ParaHox A clusters are much larger than the amphioxus cluster, although the relative spacing of genes is conserved. From top; amphioxus (*Branchiostoma floridae*), Indonesian coelacanth (*Latimeria menadoensis*), frog (*Xenopus tropicalis*), Mouse (*Mus musculus*), and human (*Homo sapiens*).

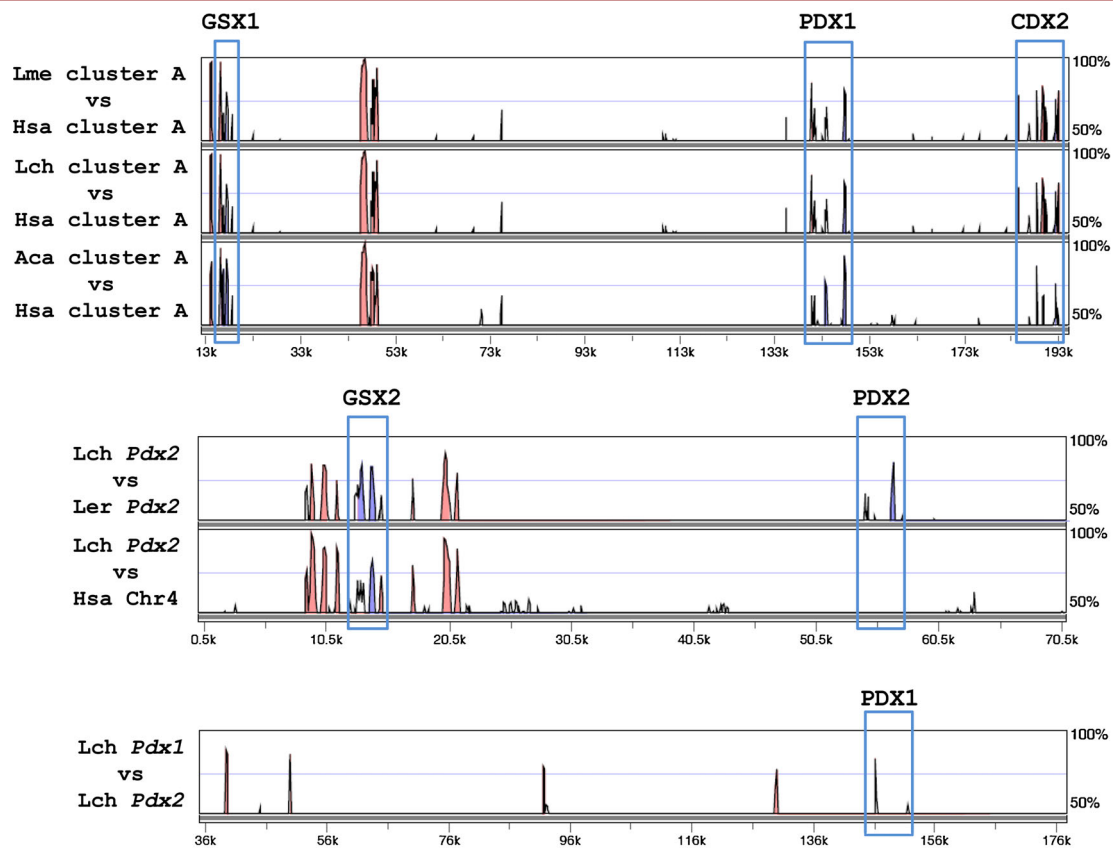


Figure 4. mVISTA alignments of vertebrate ParaHox regions. Top: Alignment of Indonesian coelacanth (Lme), African coelacanth (Lch), and Bowfin (Aca) A clusters against the human A cluster; middle—Alignment of Little skate (Ler) and human (Hsa) C clusters against the African coelacanth C cluster; bottom—Alignment of clusters A and C in African coelacanth (Lch). The conserved non-coding sequences in the bottom pane share high sequence similarity with non-coding sequences in other Coelacanth BAC clones and therefore likely represent a Coelacanth-specific repetitive element.

and maintenance of pancreatic islets (Fujitani et al., 2006)), Area IV (involved in driving gene expression specifically in β -cells (Gerrish et al., 2004)) and the E-box (involved in β -cell-specific expression of *Pdx1* (Melloul et al., 2002)). We find sequence 5' to coelacanth *Pdx1* gene corresponding to Area I (−2,761 to −2,457 in human), Area II (−2,153 to −1,923 in human), and Area III (−1,879 to −1,600 in human) (see Figs. S1–S3 for alignments). We found no evidence for Area IV (−8,656 to −8,155 in human) or the E-box (−104 in human) in the proximal promoter region.

The extent of sequence conservation between the two extant coelacanth species was also examined. Analyses of ParaHox genes for which complete gene sequence was available from both species (*Gsx1*, *Pdx1*, *Cdx2*, *Cdx4*) revealed that coding sequences were 100% conserved, whilst introns were typically >99% identical (*Gsx1* intron, 99.8; *Pdx1* intron, 99.6; *Cdx2* introns, 99.1; *Cdx4* introns, 99.8). Furthermore, a larger comparison across the entire ParaHox A cluster (from the start codon of *Gsx1* to the start of *Cdx2*, comprising over 120 kb; discounting regions with Ns in *L. chalumnae*) revealed >99% similarity. These findings are in line with those reported elsewhere and show that the ParaHox cluster is conserved at a similar level to the Hox clusters of the two coelacanth species, and slightly more highly conserved than some other genomic regions (see Amemiya et al., 2013; Table S11). This is an extremely high level of sequence similarity between the two species, considering the suggestion that they diverged 24–44 million years ago (Inoue et al., 2005). By comparison, New World monkeys (Platyrrhini) and Old World monkeys and apes (Catarrhini) diverged ~35 million years ago (Schrägo and Russo, 2003) and our analysis of sequence conservation between human and Common Marmoset (*Callithrix jacchus*) ParaHox A clusters (>150 kb) showed 66% sequence similarity.

DISCUSSION

The presence of seven ParaHox genes in coelacanths lends support to a model for ParaHox gene evolution involving whole genome duplications (WGDs) early in vertebrate ancestry followed by gene losses. It suggests that the seven ParaHox gene state found previously in cartilaginous fish (Mulley and Holland, 2010) and now also demonstrated in coelacanths was established rapidly following WGD. Analyses of the genes flanking the coelacanth ParaHox clusters (see Figs. 1 and 2) identifies several other gene families that were duplicated via WGD (*Platelet-derived growth factor receptor*, *Ligand of Numb Protein X*, *FMS-like tyrosine kinase*) and show conserved synteny to human and other vertebrate ParaHox clusters. The apparent slow rate of molecular evolution and genomic rearrangement in cartilaginous fish and coelacanths (Mulley and Holland, 2010; Wang et al., 2009) suggests that both may be excellent models for the study of events post-WGD and later genomic re-organization in the actinopterygian and sarcopterygian lineages.

Comparison of coelacanth ParaHox cluster sequences to known *Pdx1* regulatory regions from mammals (Melloul et al., 2002;

Gerrish et al., 2004; Fujitani et al., 2006) reveals that several regulatory elements are more ancient than previously thought (see alignments in Supplementary information). Our results also suggest that the loss of Area II in birds is potentially unique to that lineage (or possibly unique to reptiles, as we have also not been able to find any similarity to this region in the sequence upstream of the Green Anole (*Anolis carolinensis*) *Pdx1* gene in the AnoCar2.0 genome assembly). In Areas I–III, the highest level of sequence conservation was seen in only part of the region previously defined (Gerrish et al., 2000) and it is possible that these areas represent core enhancer regions that have been elaborated on later in tetrapod or mammalian evolution, as has been reported for lamprey conserved non-coding sequences (McEwen et al., 2009). Our findings from the coelacanths may be particularly useful for future studies of these sites, as many of the regions that are 100% conserved are around 8–10 bp in length, similar to the size of most transcription factor binding sites. Surprisingly, we were unable to find any conserved non-coding sequences between the *Pdx1* and *Pdx2* gene-containing regions of coelacanths suggesting that these genes do not share regulatory architecture. We were also unable to identify any conserved non-coding sequences between the *Pdx2*-containing regions of coelacanth and Little skate, suggesting there may have been differential loss of regulatory elements following WGD. Given the difficulty in accessing tissue samples from the CITES-listed coelacanths, it is unlikely that we will be able to discover the function of the *Pdx2* gene in this lineage. Model cartilaginous fish such as the Little skate (*L. erinacea*) or Lesser Spotted Catshark (*S. canicula*), both of which have ongoing transcriptome and whole genome sequencing projects, offer the best chance to determine the role of the *Pdx2* gene in embryos and adults.

MATERIALS AND METHODS

Isolation of Indonesian Coelacanth (*Latimeria menadoensis*) ParaHox Genes by Degenerate PCR

Homeobox-containing fragments of members of the *Gsx*, *Xlox*, and *Cdx* gene families were isolated from genomic DNA using the following degenerate PCR primers: for *Gsx* JMGsx1a: 5'-ATG YCG MGV TTY TTY YWB GT-3' (forward); JMGsx1b: 5'-GTN GAY TCN YTV ATN WTN ARG GA-3' (nested forward); *Gsx3*: 5'-TTG CCY TCY TTY TTG TGC TT-3' (reverse); *GsxS02*: 5'-CAN CKD CGR TTY TGR AAC CA-3' (nested reverse); for *Cdx* JmCdx: 5'-GGN AAR CAN MGR ACV AAR GA-3' (forward); *CdxS01*: 5'-CTRGARCTG-GARAARGARTT-3' (nested forward); *CdxS02*: 5'-NVK NVK RTT YTG RAA CCA-3' (reverse) and for *Pdx* JMXlox1c: 5'-GAC GAC AAC AAG MGN CAN AGR AC-3' (forward); *Xlox2*: 5'-CAG CTG CTV GAG CTV GAG AA-3' (nested forward); *Xlox3*: 5'-YTC CTC YTT YTT CCA CTT CAT-3' (reverse); *XS02*: 5'-GCG NCG RTT YTG GAA CCA GAT-3' (nested reverse). The resulting homeobox fragments (representing two *Gsx* genes, two *Pdx* genes, and three *Cdx* genes) were digoxigenin (DIG)-labeled and used to screen a

high coverage BAC library for Indonesian coelacanth (from the Genome Resource Centre, Benaroya Research Institute, Seattle, WA, USA (Danke et al., 2004)). Positive clones were verified by PCR prior to sequencing and were found to contain *Pdx1*-like and *Cdx2*-like gene fragments (Clones 95G3 and 161J8), a *Pdx2*-like fragment (Clone 188I4) and a *Cdx4*-like fragment (Clone 52G17). No *Cdx1*-positive clones were found during this library screening, nor were they identified using a variety of DIG-labelled homeobox probes. As previously reported (Mulley and Holland, 2010), clone 188I4 containing *Pdx2* (accession HM134895) was sequenced to 9.7× coverage using Sanger sequencing (performed at the Washington University Genome Centre, St. Louis, MI) and clones 52G17 (*Cdx4*) and 95G3 (*Pdx1*) were sequenced in the same way to 10.1× coverage and 6.0× coverage respectively. Clone 161J8 was sequenced to ~40× coverage using Roche 454 GS FLX Titanium technology (performed at the Centre for Genomic Research, University of Liverpool, UK). Genes were predicted using BLAST and GenScan and by alignment to known orthologous genes from other vertebrates. BAC clone sequences are deposited in GenBank under accession numbers KC914566–KC914567.

Phylogenetic Analysis

Orthology of genes encoded by the sequenced BAC clones was confirmed using phylogenetic analysis. Amino acid sequences were aligned using ClustalX (Larkin et al., 2007) and edited by eye to maximize contiguity of alignable sequence; maximum likelihood phylogenetic trees were constructed with PhyML (Guindon and Gascuel, 2003) using the JTT matrix and 1,000 bootstrap replicates. All nodes had >50% support and trees were rooted with amphioxus (*Branchiostoma floridae*) or fly (*Drosophila melanogaster*) sequences. The resulting trees are provided as Supplementary information. Tajima's Relative Rate Test (Tajima, '93) was conducted on amino acid sequences in MEGA version 5 (Tamura et al., 2011) using amphioxus Xlox as an outgroup. All positions containing gaps were eliminated.

Coelacanth Genome Survey and Comparative Genomics

Genomic data for the African coelacanth (*L. chalumnae*) became available in late 2011 and was used to fill gaps in our *L. menadoensis* BAC clone data. BLAST surveys using *L. menadoensis* genes were carried out against the LatChal September 2011 v64 assembly at PreEnsembl (<http://pre.ensembl.org>) and comparisons for sequence conservation were carried out using mVISTA (Frazer et al., 2004) and the AVID alignment program (Bray et al., 2003) using a 100 bp window 70% conservation level. Additional analyses of conserved non-coding sequences and pairwise alignments of *L. menadoensis* and *L. chalumnae* sequences were carried out using ClustalW (Larkin et al., 2007) implemented in BioEdit (Hall, '99). Common Marmoset (*C. jacchus*) genome sequence analyses were based on CalJac3.2.1 (Release 58, May 2010) data from Ensembl—www.ensembl.org/Callithrix_jacchus.

ACKNOWLEDGMENTS

Coelacanth genomic DNA was kindly provided by Chris Amemiya and Axel Meyer. We thank Jordi Paps for discussions and advice. The authors also wish to acknowledge the efforts and achievements of the Coelacanth Genome Sequencing Project. This work was supported by the Wellcome Trust.

LITERATURE CITED

- Amemiya CT, Powers TP, Prohaska SJ, et al. 2010. Complete HOX cluster characterization of the coelacanth provides further evidence for slow evolution of its genome. *Proc Natl Acad Sci* 107:3622–3627.
- Amemiya CT, Alföldi J, Lee AP, et al. 2013. The African coelacanth genome provides insights into tetrapod evolution. *Nature* 496:311–316.
- Arnold M, Rizzo F, Annunziata R, et al. 2006. Genetic organization and embryonic expression of the ParaHox genes in the sea urchin *S. purpuratus*: insights into the relationship between clustering and colinearity. *Dev Biol* 300:63–73.
- Boyer DF, Fujitani Y, Gannon M, et al. 2006. Complementation rescue of Pdx1 null phenotype demonstrates distinct roles of proximal and distal *cis*-regulatory sequences in pancreatic and duodenal expression. *Dev Biol* 298:616–6631.
- Bray N, Dubchak I, Pachter L. 2003. AVID: a global alignment program. *Genome Res* 13:97–102.
- Brooke NM, Garcia-Fernández J, Holland PWH. 1998. The ParaHox gene cluster is an evolutionary sister of the Hox gene cluster. *Nature* 392:920–922.
- Cockburn BN, Bermanno G, Boodram LLG, et al. 2004. Insulin promoter factor-1 mutations and diabetes in Trinidad: identification of a novel diabetes-associated mutation (E224K) in an Indo-Trinidadian family. *J Clin Endocrinol Metab* 89:971–978.
- Danke J, Miyake T, Powers T, et al. 2004. Genome resource for the Indonesian coelacanth, *Latimeria menadoensis*. *J Exp Zool A Comp Exp Biol* 301A:228–234.
- Ferrier DEK, Minguillón C. 2003. Evolution of the Hox/ParaHox gene clusters. *Int J Dev Biol* 47:605–612.
- Ferrier DE, Dewar K, Cook A, et al. 2005. The chordate ParaHox cluster. *Curr Biol* 15:R820.
- Frazer KA, Pachter L, Poliakov A, Rubin EM, Dubchak I. 2004. VISTA: computational tools for comparative genomics. *Nucleic Acids Res* 32(Suppl 2):W273–W279.
- Fujitani Y, Fujitani S, Boyer DF, et al. 2006. Targeted deletion of a *cis*-regulatory region reveals differential gene dosage requirements for Pdx1 in foregut organ differentiation and pancreas formation. *Genes Dev* 20:253–266.
- Gerrish K, Gannon M, Shih D, et al. 2000. Pancreatic B-cell-specific transcription of the Pdx-1 gene. *J Biol Chem* 275:3485–3492.
- Gerrish K, Van Velkinburgh JC, Stein R. 2004. Conserved transcriptional regulatory domains of the pdx-1 gene. *Mol Endocrinol* 18:533–5548.

- Guindon S, Gascuel O. 2003. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst Biol* 52:696–704.
- Hall TA. 1999. BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. *Nucleic Acids Res Symp Ser* 41:95–98.
- Hani EH, Stoffers DA, Chèvre JC, et al. 1999. Defective mutations in the insulin promoter factor-1 (IPF-1) gene in late-onset type 2 diabetes mellitus. *J Clin Invest* 104:R41.
- Holland PWH. 2001. Beyond the Hox: how widespread is homeobox gene clustering? *J Anat* 199:13–23.
- Hsieh-Li HM, Witte DP, Szucsik JC, et al. 1995. Gsh-2, a murine homeobox gene expressed in the developing brain. *Mech Dev* 50:177–186.
- Hui JHL, Holland PWH, Ferrier DEK. 2008. Do cnidarians have a ParaHox cluster? Analysis of synteny around a *Nematostella* homeobox gene cluster. *Evol Dev* 10:725–730.
- Hui J, Raible F, Korchagina N, et al. 2009. Features of the ancestral bilaterian inferred from *Platynereis dumerilii* ParaHox genes. *BMC Biol* 7:43.
- Inoue JG, Miya M, Venkatesh B, Nishida M. 2005. The mitochondrial genome of Indonesian coelacanth *Latimeria menadoensis* (Sarcopterygii: Coelacanthiformes) and divergence time estimation between the two coelacanths. *Gene* 349:227–235.
- Jonsson J, Carlsson L, Edlund T, Edlund H. 1994. Insulin-promoter-factor 1 is required for pancreas development in mice. *Nature* 371:606–609.
- Kulakova MA, Cook CE, Andreeva TF. 2008. ParaHox gene expression in larval and postlarval development of the polychaete *Nereis virens* (Annelida, Lophotrochozoa). *BMC Dev Biol* 8:61.
- Larkin MA, Blackshields G, Brown NP, et al. 2007. Clustal W and Clustal X version 2.0. *Bioinformatics* 23:2947–2948.
- Macfarlane WM, Frayling TM, Ellard S, et al. 1999. Missense mutations in the insulin promoter factor-1 gene predispose to type 2 diabetes. *J Clin Invest* 104:R33–R39.
- McEwen GK, Goode DK, Parker HJ, et al. 2009. Early evolution of conserved regulatory sequences associated with development in vertebrates. *PLoS Genet* 5:e1000762.
- Melloul D, Marshak S, Cerasi E. 2002. Regulation of pdx-1 gene expression. *Diabetes* 51(Suppl 3):S320–S325.
- Mendivil Ramos O, Barker D, Ferrier DEK. 2012. Ghost loci imply Hox and ParaHox existence in the last common ancestor of animals. *Curr Biol* 22:1951–1956.
- Miyatsuka T, Matsuoka T, Shiraiwa T, et al. 2007. Ptf1a and RBP-J cooperate in activating Pdx1 gene expression through binding to Area III. *Biochem Biophys Res Commun* 362:905–909.
- Moreno E, Morata G. 1999. Caudal is the Hox gene that specifies the most posterior *Drosophila* segment. *Nature* 400:873–877.
- Mulley JF, Holland PWH. 2010. Parallel retention of Pdx2 genes in cartilaginous fish and coelacanths. *Mol Biol Evol* 27:2386–2391.
- Mulley JF, Chiu C, Holland PWH. 2006. Breakup of a homeobox cluster after genome duplication in teleosts. *Proc Natl Acad Sci* 103:10369–10372.
- Noonan JP, Grimwood J, Danke J, et al. 2004. Coelacanth genome sequence reveals the evolutionary history of vertebrate genes. *Genome Res* 14:2397–2405.
- Offield MF, Jetton TL, Labosky PA, et al. 1996. PDX-1 is required for pancreatic outgrowth and differentiation of the rostral duodenum. *Development* 122:983–995.
- Ohlsson H, Karlsson K, Edlund T. 1993. IPF1, a homeodomain-containing transactivator of the insulin gene. *EMBO J* 12:4251–4259.
- Powers TP, Amemiya CT. 2004. Evidence for a Hox14 paralog group in vertebrates. *Curr Biol* 14:R183–R184.
- Reece-Hoyes JS, Keenan ID, Isaacs HV. 2002. Cloning and expression of the Cdx family from the frog *Xenopus tropicalis*. *Dev Dyn* 223:134–140.
- Rosanas-Urgell A, Marfany G, Garcia-Fernandez J. 2005. Pdx1-related homeodomain transcription factors are distinctly expressed in mouse adult pancreatic islets. *Mol Cell Endocrinol* 237:59–66.
- Samadi L, Steiner G. 2010. Conservation of ParaHox genes' function in patterning of the digestive tract of the marine gastropod *Gibbula varia*. *BMC Dev Biol* 10:74.
- Schrägo CG, Russo CAM. 2003. Timing the origin of New World monkeys. *Mol Biol Evol* 20:1620–1625.
- Schwitzgebel VM, Mamin A, Brun T, et al. 2003. Agenesis of human pancreas due to decreased half-life of insulin promoter factor 1. *J Clin Endocrinol Metab* 88:4398–4406.
- Stoffers DA, Zinkin NT, Stanojevic V, Clarke WL, Habener JF. 1997. Pancreatic agenesis attributable to a single nucleotide deletion in the human IPF1 gene coding sequence. *Nat Genet* 15:106–110.
- Tajima F. 1993. Simple methods for testing the molecular evolutionary clock hypothesis. *Genetics* 135:599–607.
- Tamura K, Peterson D, Peterson N, et al. 2011. MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Mol Biol Evol* 28:2731–2739.
- Valerius MT, Li H, Stock JL, et al. 1995. Gsh1: a novel murine homeobox gene expressed in the central nervous system. *Dev Dyn* 203:337–351.
- Wang J, Lee AP, Kodzius R, Brenner S, Venkatesh B. 2009. Large number of ultraconserved elements were already present in the jawed vertebrate ancestor. *Mol Biol Evol* 26:487–490.
- Wiebe PO, Kormish JD, Roper VT, et al. 2007. Ptf1a binds to and activates area III, a highly conserved region of the Pdx1 promoter that mediates early pancreas-wide Pdx1 expression. *Mol Cell Biol* 27:4093–4104.

SUPPORTING INFORMATION

Additional Supporting Information may be found in the online version of this article at the publisher's web-site.